DO EMBODIED AI MODELS HAVE PHYSICAL INTUITION? AN EGOCENTRIC PROBING BENCHMARK

Anonymous authors
Paper under double-blind review

ABSTRACT

Estimating contact forces from egocentric observations requires embodied models to map visual appearance and hand motion to continuous physical quantities. These estimates depend on how object geometry, deformation, and load evolve during interaction. Occlusion and visual–physical mismatch complicate this mapping across materials and fill states. We introduce FEEL (Force Estimation from Egocentric Learning), a controlled benchmark of 3 069 vision–touch–pose episodes in ten object and fill-state categories, and EgoForce, which uses a GRU for visual–state aggregation and two softly weighted residual experts to refine a shared Laplace force head. We evaluate transformer regression policies, flow-matching vision–language–action models, generalist multimodal policies, and vision-foundation encoders on current-force prediction from RGB and hand state. The 36.18 M-parameter EgoForce achieves test contact MAE of 7.3386 N, contact R^2=0.1965, contact Pearson r=0.3022, and pooled F1 of 0.8395, with OOD contact r=0.4067 and pooled F1 of 0.8934 on empty milk containers. Project page: `http://submission.anonymoussubmission.com/`

1 INTRODUCTION

Humans possess physical intuition prior to any physical interaction. Cognitive science shows that human manipulation relies on internal forward models driven by visual cues (Wolpert et al., 1995; Wolpert & Flanagan, 2001). For instance, when a person sees a cup filled with water, they anticipate the weight and apply a larger initial grasping force. When encountering a visual-physical mismatch (for example, picking up an opaque Coke can that appears full but is actually empty), humans exhibit a predictable force overshoot before rapidly correcting via tactile feedback. This raises a key question: whether embodied AI models learn these physical predictions and misjudgments.

Recent work has produced embodied foundation models and visuomotor policies (e.g., ACT (Zhao et al., 2023), Pi0, GR00T) (Zitkovich et al., 2023; Kim et al., 2024) that generalise across diverse semantic tasks. However, "semantic intelligence" does not equate to "physical intelligence." A model may follow the instruction "pick up the egg," but if it lacks the intuition of *how much force* to apply, the task fails. We therefore ask: *can current embodied AI models learn the same physical intuition, and under what circumstances do they struggle?*

Concurrently, egocentric (first-person) data collection using wearable devices such as Meta Aria (Engel et al., 2023) has emerged as a scalable paradigm to overcome the teleoperation bottleneck, mirroring internet-scale datasets (Grauman et al., 2022). Recent egocentric tactile datasets such as OpenTouch (Song et al., 2025) demonstrate that in-the-wild vision-touch-pose capture is feasible at scale. However, existing benchmarks emphasise broad perception tasks such as cross-modal retrieval, grasp classification, and contact understanding. Controlled comparisons across fill levels and visually ambiguous loads complement these perception tasks by targeting continuous force estimation at the thumb and index finger throughout the recorded grasping sequence.

To bridge this gap, we introduce **FEEL** (Force Estimation from Egocentric Learning), an offline representation benchmark and reproducible data engine for *controlled force-intuition probes*. The benchmark comprises 3 069 episodes of synchronised vision–touch–pose data across ten categories defined by object type and fill state, organised within the Fragile, Deformable, and Variable-Load taxonomy. To analyse the force-learning problem, we introduce explicit physical inductive biases for force estimation and systematically evaluate them on FEEL. Our contributions are:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

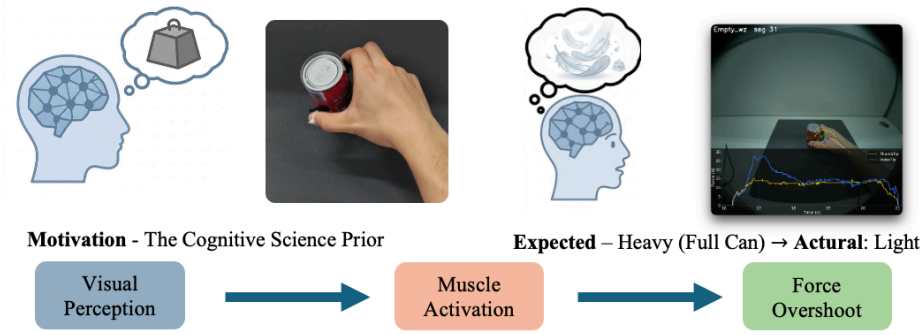


Figure 1: The cognitive-science prior FEEL probes. A familiar Coke can primes a heavy-load expectation; when the can is actually empty, the predicted force *overshoots* before tactile feedback corrects it (**Visual Perception** → **Muscle Activation** → **Force Overshoot**). FEEL operationalises this prior as a controlled benchmark for diagnosing physical intuition in embodied AI models.

1. **The FEEL Benchmark and Probing Protocol:** We propose an offline representation benchmark with an F/D/V (Fragile/Deformable/Variable-Load) taxonomy. Our protocol captures controlled *Boundary Samples* (e.g., deliberate physical failures) and *Visual-Physical Mismatch Samples* (e.g., deceptive weights) to probe force estimation when visual priors are unreliable. We adopt contact MAE, contact R^2, episode-averaged contact Pearson r, and pooled contact-detection F1 as primary metrics for evaluating force estimation.
2. **Critical Analysis of Embodied AI Models:** We evaluate eight representative baselines (ACT, WALL-X, Pi0, Pi0.5, X-VLA, SmolVLA, OpenTouch, GR00T) on FEEL. The lightweight ACT baseline achieves higher test contact R^2 than Pi0, Pi0.5, and X-VLA, showing that larger pre-trained backbones do not consistently yield better force estimation on FEEL across models ranging from tens of millions to several billion parameters.
3. **Physical Inductive Biases for Force Estimation:** We introduce **EgoForce**, which combines visual–state temporal modeling with a shared Laplace force head and two residual experts that refine force means. In the main benchmark, 36.18 M-parameter EgoForce achieves the best test contact MAE, contact R^2, and contact Pearson r. Its shared-head variant leads in test pooled F1; EgoForce leads on all four OOD metrics.

All datasets, benchmark evaluation code, and hardware blueprints for the data engine will be publicly released under the MIT/CC BY 4.0 license.

2 Related Work

Tactile Demonstrations and Scalable Egocentric Collection. A line of work shows that force-aware human demonstrations can train contact-rich manipulation policies more efficiently than teleoperation: MimicTouch (Yu et al., 2024) achieves higher collection throughput via GelSight tactile capture; Feel the Force (Adeniji et al., 2025) uses a force glove with a Transformer policy; ViTaMIn (Liu et al., 2025), TAMEn (Wu et al., 2026a), HTD (Niu et al., 2026), and TF-Gripper (Kang et al., 2026) scale tactile demonstrations to bimanual and humanoid settings. In parallel, scalable egocentric collection systems (UMI (Chi et al., 2024), DexCap (Wang et al., 2024), EgoMimic (Kareer et al., 2025), Ψ_0 (Wei et al., 2026)) support scalable collection of first-person manipulation demonstrations for learning visuomotor policies from recorded human interactions.

Visuo-Tactile Fusion and Force Estimation from Vision. Cognitive science establishes that humans use learned visual-physical priors to anticipate contact forces before any touch occurs (Wolpert et al., 1995; Wolpert & Flanagan, 2001). Computationally, PressureVision (Grady et al., 2022) demonstrates that RGB images naturally encode contact pressure without any tactile sensor, and PressureVision++ (Grady et al., 2024) extends this to in-the-wild fingertip-level prediction. EgoTouch (Mollyn & Harrison, 2024) shows that egocentric AR cameras resolve fine-grained contact

events, and EgoPressure (Zhao et al., 2025) pairs egocentric imagery with pressure-sensitive surface measurements and multi-view hand-pose annotations for joint pose and pressure estimation—the most direct predecessor to FEEL. Across simulation and real-robot settings, visuo-tactile fusion consistently improves manipulation generalization (Sferrazza et al., 2024; Guzey et al., 2024; Zheng et al., 2026), while OpenTouch (Song et al., 2025) provides the first large-scale in-the-wild dataset pairing egocentric observations with full-hand tactile recordings of object interaction.

VLA Foundation Models and Emerging Force-Aware Systems. State-of-the-art VLA models—Pi0 (Black et al., 2024), Pi0.5 (Black et al., 2025), X-VLA (Zheng et al., 2025), SmolVLA (Shukor et al., 2025), and WALL-X (Zhai et al., 2025)—achieve strong results on semantic manipulation benchmarks through large-scale generative pre-training (Zitkovich et al., 2023; Kim et al., 2024), with generative policy architectures enabling continuous action distributions (Chi et al., 2023; Black et al., 2024). Force-aware systems incorporate force information through control, distillation, and retrieval: ForceVLA2 (Li et al., 2026) integrates hybrid force-position control with cross-scale MoE routing, HapticVLA (Gubernatorov et al., 2026) distills tactile knowledge into vision-only deployment, and Exp-Force (Shang et al., 2026) retrieves grasp-force experience from VLMs.

3 THE FEEL BENCHMARK

This section details the **FEEL Benchmark** and its egocentric visuo-tactile collection setup. FEEL evaluates force estimation across material properties, fill levels, and visually ambiguous load conditions. The collection contains **449.5 minutes** (≈7.5 hours) of raw recordings from **four demonstrators**; processing yields **85 177 frames**. The **10 benchmark categories**, defined by object type and fill state, are potato chips (Chip), jelly, sponges, empty bottles, bottles filled to 25%, 50%, and 75% (P25, P50, and P75 Bottle), full bottles, milk-filled containers (Filled Milk), and empty milk containers (Empty Milk). F/D/V describes their physical properties.

3.1 REPRODUCIBLE VISUO-TACTILE FORCE-PROBE ENGINE

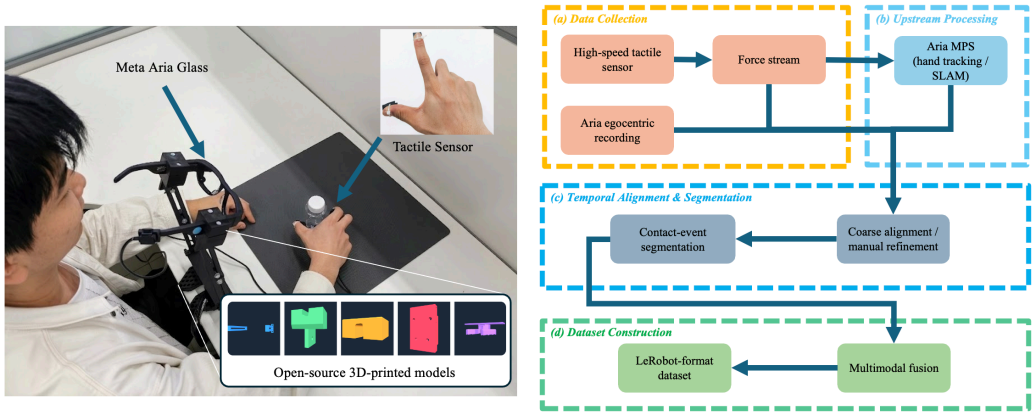
A major bottleneck in force-intuition learning is the lack of reproducible protocols for synchronized visual, pose, and tactile measurements under controlled physical counterfactuals. We therefore release a low-cost data collection setup that combines three components. The **hardware blueprint and software suite** provides full 3D-printable CAD models, a complete Bill of Materials integrating Meta Aria glasses with the **Paxini PX6AX-GEN3-DP-S2015-Elite** tactile sensor for high-fidelity fingertip force/deformation feedback, plus collection and conversion scripts for reproducible early-contact force curves. The **spatio-temporal alignment pipeline** offers a browser-based UI for timestamp alignment and event-based offset refinement between Aria visual/IMU streams and tactile readings, together with manual and automatic spatial calibration algorithms that register the tactile sensor frame to the Aria egocentric coordinate system. Finally, **MPS-based hand tracking** provides the fingertip positions used to construct the 7-D hand-state input without external motion capture.

3.2 F/D/V PHYSICAL TAXONOMY

We organise the collection along three physical axes (Figure 3). The **Fragile (F)** axis includes *potato chips*, with boundary demonstrations recording the transition from intact grasping to material failure. The **Deformable (D)** axis covers *plastic bottles, jelly cups, and sponges*, which exhibit visible deformation during contact. The **Variable-Load (V)** axis includes bottles at five fill levels and filled or empty milk containers, combining visible fill-level changes with load variation concealed by opaque packaging. These axes describe overlapping physical properties: a filled plastic bottle can be both deformable and variable-load.

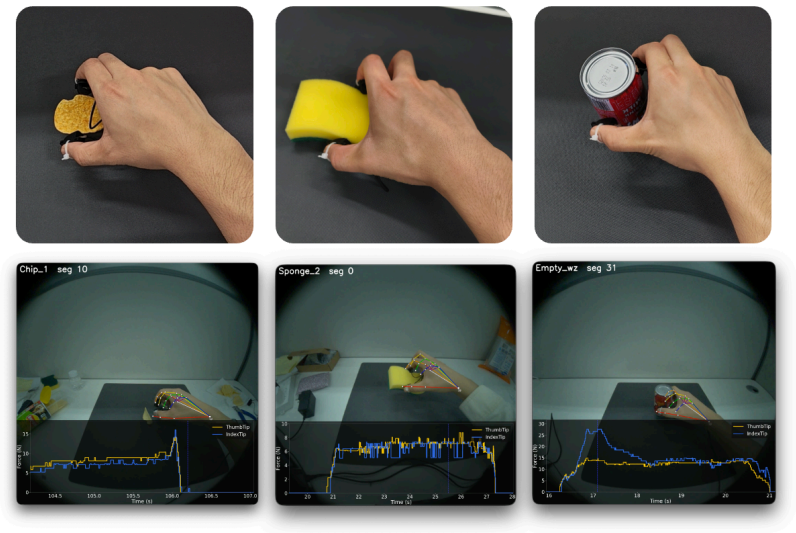
3.3 STANDARDIZED PROTOCOL WITH BOUNDARY AND MISMATCH SAMPLES

The standard collection sequence has five stages: **Approach, Grasp, Lift, Carry, and Place**. The protocol additionally includes three types of physical probes: (i) **Boundary Samples**, which record material failure during grasping, such as a potato chip breaking under compression; (ii) **Water-Level Samples**, which vary bottle fill levels across 0%, 25%, 50%, 75%, and 100% to study visible load



(a) Hardware & Collection (b) Processing Pipeline

Figure 2: The FEEL data engine. (a) Hardware: Meta Aria glasses paired with Paxini tactile sensors on thumb and index finger via 3D-printed brackets. (b) Processing pipeline: synchronized streams flow through MPS-based tracking, temporal alignment, contact segmentation, and multimodal fusion into LeRobot-format episodes.



(a) Fragile (F) (b) Deformable (D) (C) Variable-Load (V)

Figure 3: Representative episodes illustrating the three F/D/V physical properties. **Top:** third-person view. **Bottom:** egocentric (Aria) view with per-finger force trace. (a) **Fragile**: a potato chip shatters under thumb pressure. (b) **Deformable**: a sponge undergoes elastic compression. (c) **Variable-Load**: an empty opaque bottle whose appearance does not betray its true mass.

cues and initial grasping force; and (iii) **Visual-Physical Mismatch Samples**, which include empty opaque containers whose appearance conceals their contents.

Task and supervision. Each synchronized sample pairs egocentric RGB and a 7-D hand state with thumb and index-finger force measurements. Models estimate the current force on each finger from visual and hand-state observations, using tactile measurements as supervision. Continuous recordings from approach through placement support comparison of force magnitude and temporal variation across changes in hand configuration, object deformation, and load during each interaction.

Demonstrators. Four demonstrators provided data; releases follow the LeRobot dataset standard.

4 PHYSICS-AWARE FORCE PREDICTION ARCHITECTURE

EgoForce builds on the encoder–decoder architecture of ACT (Zhao et al., 2023). It first aggregates visual–state observations into a temporal representation, then predicts a shared force distribution whose mean is refined by input-dependent residual experts (Figure 4).

4.1 MULTI-MODAL ENCODING AND TEMPORAL MODELING

Contact force varies with approach motion, contact-onset transients, and sustained deformation. A GRU aggregates a window of n_{obs}=4 frames to model local temporal dependencies and predicts the current force at the final observation timestep in each window. Since the hand frequently occludes the object during grasping, we additionally encode a 7-dimensional hand state—thumb fingertip position (x, y, z), index fingertip position (x, y, z), and thumb–index distance d_{ti}, derived from MPS hand tracking—to provide explicit grasp geometry alongside the visual observations.

4.2 LAPLACE FORCE HEAD WITH SCALE CLAMP

The temporal representation $\mathbf{h}_t$ feeds two finger-specific heads, each predicting a base mean $\mu_{t,f}^{\text{base}}$ and a log scale $s_{t,f}$. For a quantile-normalized force target $y_{t,f}$, the clamped-scale Laplace objective combines a scale-weighted absolute error with a log-scale penalty, averaged over frames and fingers:

$$\mathcal{L}_{\text{force}} = \mathbb{E}_{t,f}\left[\frac{|y_{t,f} - \mu_{t,f}|}{\exp(\bar{s}_{t,f})} + \bar{s}_{t,f}\right], \qquad \bar{s}_{t,f} = \text{clip}(s_{t,f}, -5, 2), \tag{1}$$

where $\mu_{t,f}$ includes the residual correction defined below. The same clamped log scale determines both the error weight and the additive penalty, with the force scale bounded between e^{-5} and e^{2}. An auxiliary contact head on $\mathbf{h}_t$ adds binary cross-entropy with weight 0.1.

4.3 RESIDUAL MIXTURE-OF-EXPERTS FORCE HEAD

Different objects and fill states produce different force responses within a shared visual–temporal representation. EgoForce models these variations at the force output. A linear router computes $\boldsymbol{\pi}_t = \text{softmax}(W_r \mathbf{h}_t + \mathbf{b}_r)$, and two MLP experts predict corrections to the two-finger force mean:

$$\boldsymbol{\mu}_t = \boldsymbol{\mu}_t^{\text{base}} + \sum_{k=1}^{2} \pi_{t,k}\, \Delta\boldsymbol{\mu}_k(\mathbf{h}_t). \tag{2}$$

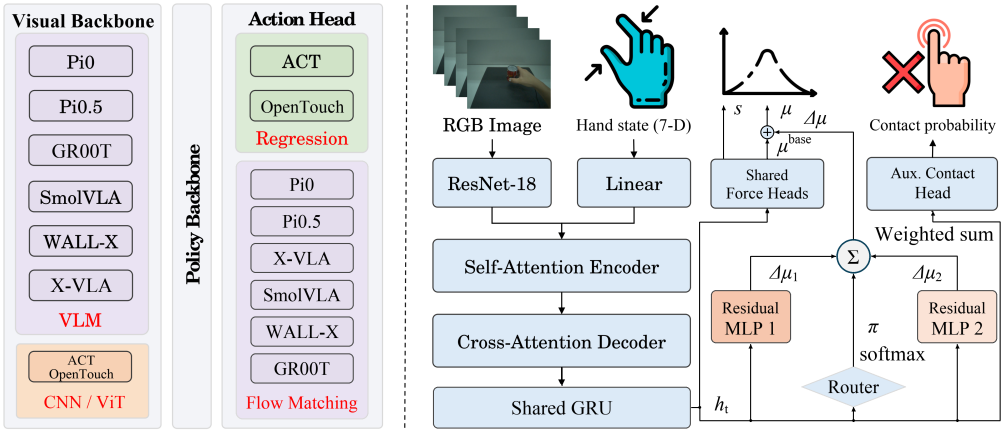
Each expert has dimensions $512 \to 128 \to 2$ with a GELU activation. Both experts contribute to every prediction, with weights determined by the final GRU hidden state. The finger-specific scale outputs remain in the shared Laplace heads. This separates temporal aggregation from the input-dependent correction of force means.

The model is trained jointly for 10 000 steps from an ImageNet-initialized ResNet-18 and newly initialized task modules. Each residual expert uses an independently initialized hidden layer and a zero-initialized output projection, giving zero mean correction at initialization. The residual experts and router add 132 870 parameters to the 36 044 231-parameter shared model, yielding 36 177 101 parameters in total, an increase of approximately 0.37% over the shared-head model.

5 EXPERIMENTS

Our experiments address four questions: **(Q1)** How does force-estimation performance vary across model architectures and scales? **(Q2)** How well do models generalize to unseen opaque-container conditions? **(Q3)** How do force-command errors affect tracking accuracy and manipulation success? **(Q4)** How accurately can tactile feedback track prescribed forces on a real gripper?

For force estimation, we compare EgoForce with ACT, a frozen-DINOv2 regression adaptation of OpenTouch, and six flow-matching VLAs on **FEEL**. All models use egocentric RGB and a 7-D hand state to predict the current two-finger forces, supervised by synchronized tactile measurements. We use seed 1000 and 10 000-step checkpoints (implementation and evaluation in Appendix B).



(a) Baseline architectures (b) Our architecture

Figure 4: EgoForce architecture. (a) Baseline backbones and force-output heads. (b) RGB and 7-D hand state are fused by the visual–state encoder–decoder and aggregated by a shared GRU. Two softly weighted residual MLP experts refine only the force means. Log scales come directly from the shared force heads; an auxiliary head predicts contact.

Dataset Splits. We use nine categories for training and in-distribution evaluation and reserve empty milk containers (*empty_want_milk*) for out-of-distribution (OOD) evaluation. The splits contain 1 937 training episodes (55 525 frames), 871 in-distribution test episodes (22 630 frames), and 261 OOD episodes (7 022 frames). We use stratified episode-level training/test partitioning (approximately 80/20) for a portion of the recordings and source-session partitioning for the remainder. All frames of an episode belong to the same split.

Evaluation Metrics. We report four primary metrics. Let $g_{e,t,f}$ be the measured force for episode e, frame t, and finger f, and let $p_{e,t,f} = \max(0, \hat{F}_{e,t,f})$ be the nonnegative prediction. Ground-truth contact is $g_{e,t,f} > 0$ N and predicted contact is $p_{e,t,f} > 0$ N. **Contact MAE** (N, ↓) averages absolute errors over ground-truth contact frames within each episode and finger, then over valid fingers within an episode, and finally equally over episodes; missed-contact predictions remain included. **Contact** R^2 (↑) is computed for each finger after pooling its ground-truth contact frames, as $1 - \sum(p - g)^2 / \sum(g - \bar{g})^2$, and then averaged over the two fingers; $\bar{g}$ is that finger's mean on these contact frames. **Contact Pearson** r (↑) is computed per episode and finger on ground-truth contact frames, requiring at least five such frames and ground-truth standard deviation at least 0.1 N. Valid fingers are averaged within each episode, followed by an equal average over eligible episodes. **Pooled contact F1** (↑) is $2\mathrm{TP}/(2\mathrm{TP} + \mathrm{FP} + \mathrm{FN})$, with counts pooled over all frames and both fingers, including noncontact frames, where positive predictions are counted as false positives.

5.1 Main Benchmark: Does Scale Imply Physical Intuition?

To answer **Q1**, we compare EgoForce with eight baselines under a common force-prediction interface (Table 1). EgoForce achieves the lowest contact MAE (7.3386 N), highest contact R^2 (0.1965), and highest contact Pearson r (0.3022). Its MAE is 3.58% lower than the strongest external baseline, GR00T at chunk size 1 (7.6107 N). The shared-head variant achieves the highest pooled F1 (0.8600), followed by EgoForce with residual experts (0.8395). Among external baselines at chunk size 1, ACT's contact R^2 (0.0670) exceeds Pi0's (-0.1654), while GR00T has lower MAE than ACT (7.6107 vs. 8.4400 N). Larger backbones therefore do not consistently yield better force estimates on FEEL; ACT and GR00T rank differently under contact MAE and contact R^2.

Figure 5 compares contact MAE across materials. EgoForce has the lowest error on Empty Bottle and Chip among the six displayed models. It improves on its shared-head variant in four of nine categories. Figure 6 presents summed two-finger force trajectories on three test episodes, illustrating differences in magnitude, onset, and temporal variation.

Table 1: Benchmark results on the FEEL test split (22 630 frames, 9 categories), seed 1000, contact threshold 0 N. Contact MAE and r use episode averaging; contact R^2 pools contact frames per finger before averaging fingers; F1 pools counts over all frames and both fingers. Paired values denote chunk 1/default (ACT: 100; Pi0/Pi0.5/SmolVLA: 50; X-VLA: 30; WALL-X: 32; GR00T: 16). EgoForce parameter counts are measured from the implemented models; baseline counts follow their reported totals. Best per metric in **bold**, second best underlined.

Model	Params	Contact MAE↓	Contact R^2 ↑	Contact r ↑	Pooled F1↑
EgoForce (MoE)	36.18 M	**7.3386**	**0.1965**	**0.3022**	0.8395
EgoForce w/o MoE	36.04 M	7.3905	0.1941	0.2963	**0.8600**
ACT	34 M	8.4400/8.4349	0.0670/0.0905	0.1933/0.2022	0.6492/0.8209
WALL-X	4.2 B	7.8986/8.8121	0.0860/-0.0515	0.1226/0.0676	0.8322/0.8245
OpenTouch	87 M	7.7258	0.1044	0.1807	0.8110
GR00T	∼3 B	7.6107/7.8585	0.0157/-0.0393	0.2286/0.1663	0.8310/0.8326
Pi0.5	∼3.6 B	9.1171/10.1447	-0.1236/-0.2900	0.1286/0.0695	0.8318/0.8332
Pi0	∼3.5 B	9.5923/9.8835	-0.1654/-0.2325	0.1102/0.0862	0.8124/0.8295
SmolVLA	460 M	9.4473/9.7020	-0.1834/-0.2393	0.1017/0.0908	0.7882/0.7840
X-VLA	880 M	11.4716/9.7409	-0.5462/-0.3939	0.0415/-0.0463	0.8217/0.8246

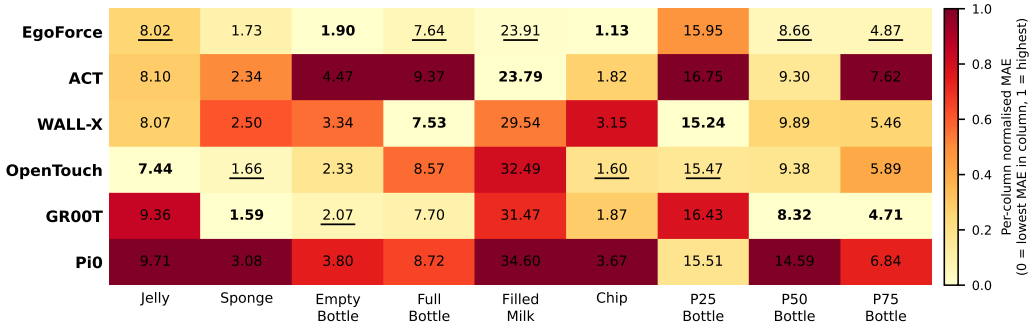


Figure 5: Per-material contact MAE on the FEEL test split, seed 1000 and chunk size 1. Errors are averaged over valid fingers within episodes and then equally over episodes in each category. Colors are normalized per column; values are in N. **Bold**: best per column; underlined: second best.

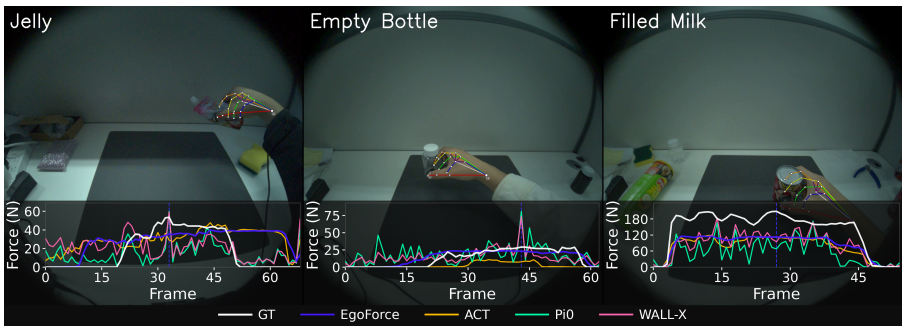


Figure 6: Force-prediction curves on Jelly, Empty Bottle, and Filled Milk test episodes (indices 0, 37, and 49). White: ground-truth force summed over both fingers; colored: model predictions from the main-benchmark 10 000-step checkpoints, seed 1000 and chunk size 1. Dashed lines link the displayed frames and hand poses to the force trajectories.

We further examine the architectural choices in **Q1** through ablations (Appendix C).

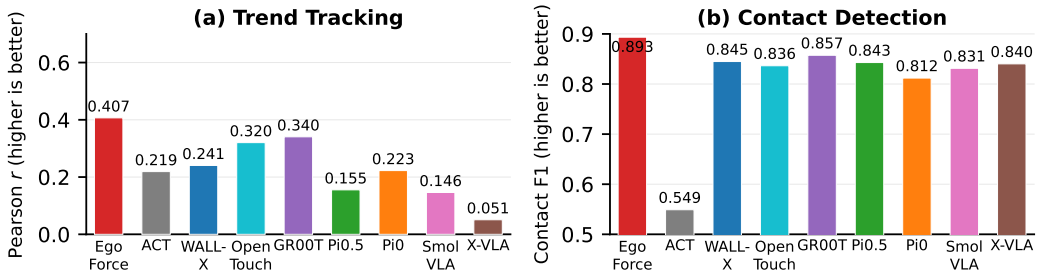


Figure 7: OOD evaluation on empty milk containers (261 episodes, 7 022 frames), seed 1000 and chunk size 1. **(a)** Episode-averaged contact Pearson r. **(b)** Pooled contact-detection F1.

5.2 OOD Generalization: Opaque Containers

To answer **Q2**, we test on held-out empty milk containers across multiple demonstrators and four source sessions (261 episodes, 7 022 frames). EgoForce achieves contact MAE of 4.3963 N, contact $R^2 = -0.7672$, contact $r = 0.4067$, and pooled F1 of 0.8934, leading all four metrics in the main benchmark. At chunk size 1, GR00T achieves MAE of 4.8675 N, $R^2 = -1.0332$, contact $r = 0.3405$, and F1 of 0.8574. X-VLA at its default chunk size 30 has the best external-baseline MAE (4.5071 N) and R^2 (-0.9889), alongside negative contact r (-0.0342). EgoForce has lower magnitude error and stronger within-episode correlation than both configurations. Relative to the shared-head variant, its MAE decreases from 6.9350 to 4.3963 N, a 36.61% reduction, while all three remaining metrics improve. Figure 7 compares EgoForce and all eight baselines at chunk size 1 using episode-averaged contact Pearson r in panel (a) and pooled contact-detection F1 in panel (b).

5.3 Force-Action Perturbations in Simulation

To answer **Q3**, we perturb the bilateral force-action channels of Tabero-VTLA (Wu et al., 2026b) on the Milk task in Isaac Sim, pairing initial states and sampling noise across perturbation levels. Across 150 trials per level, success falls from 84.7% at 0 N injected MAE to 53.3% at 4 N, 28.0% at 6 N, and 0% at 16–20 N, while force-tracking error increases (Figure 8; setup in Appendix E).

5.4 Real-Gripper Force Tracking

To answer **Q4**, we adapt the controller of Piga & Natale (2023) to track prescribed forces on a gripper using the mean feedback from electronic skin on both fingers. Holding a tissue package, the gripper tracks targets stepped from 1 to 5 N and back (Figure 9). Feedback MAE is 0.0487 N over all nine intervals, including transitions, and 0.0207 N over their final 10 s windows. Appendix D details the protocol and per-stage results.

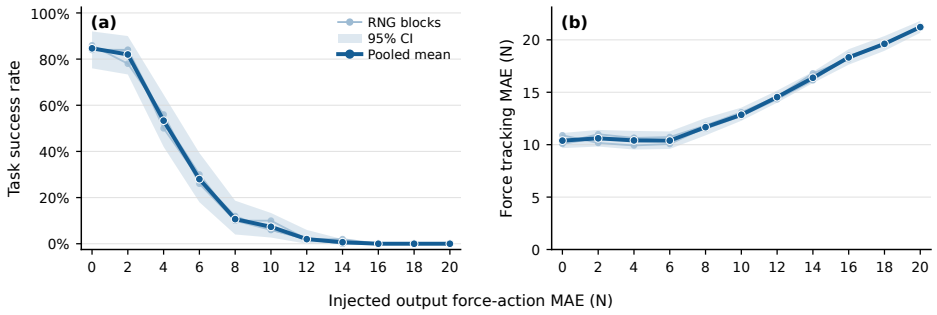
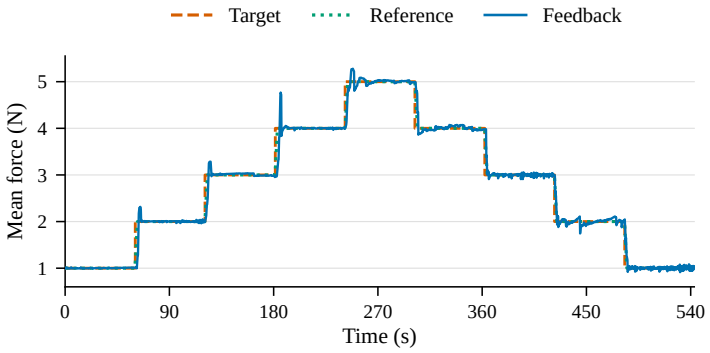


Figure 8: Controlled force-action perturbations in the simulated Milk task (150 trials per level). **(a)** Task success. **(b)** Force-tracking MAE. Shading indicates 95% confidence intervals.



(a) Gripper setup

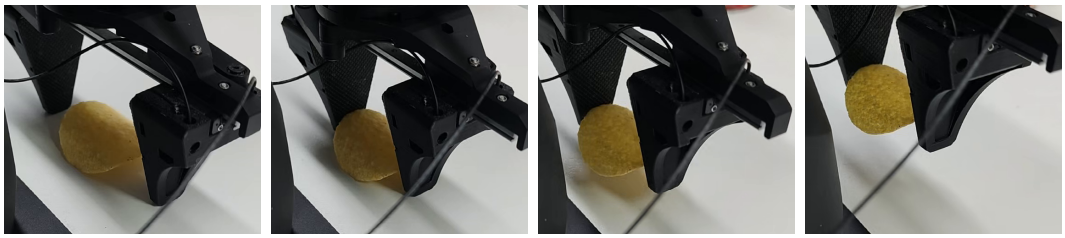


(b) Mean force tracking

Figure 9: Real-gripper force tracking on a tissue package. Targets span 1–5 N; the reference changes at 0.5 N/s. Feedback averages the two tactile readings; time starts after initial grasp establishment.

For the final 10 s windows across all nine force intervals, tracking RMSE is 0.0291 N and 92.7% of samples lie within ± 0.05 N of the prescribed force, covering targets from 1 to 5 N.

The gripper also lifts a potato chip without visible fracture (Figure 10).



(a) Before grasping (b) Grasping (c) Lifting (d) Suspended

Figure 10: Grasping and lifting a potato chip with an electronic-skin gripper. The sequence shows grasp establishment, lifting, and suspended holding without visible fracture.

6 Conclusion and Limitations

FEEL establishes a controlled benchmark for learning continuous fingertip forces from egocentric RGB and hand motion, and EgoForce combines shared temporal representations with a residual mixture-of-experts force head. The 36.18 M-parameter model achieves the best test contact MAE, contact R^2, and contact Pearson r in the main benchmark, together with the best results on all four OOD metrics; comparison with the shared-head variant identifies the benefits and metric-specific effects of the residual experts. Controlled force-command perturbations in simulation reduce manipulation success as errors increase, while tactile feedback on the real gripper tracks prescribed 1–5 N targets with a final-window MAE of 0.0207 N. Future work will extend force estimation across demonstrators and embodiments and connect the learned estimates to closed-loop manipulation.

Limitations & Broader Impact. The benchmark covers two scalar fingertip forces from four demonstrators, with seed-1000 model comparisons and an OOD split of empty milk containers from four source sessions. Held-out-person evaluation, repeated training seeds, shear forces, and torques provide natural extensions of this setting. The collection tools and evaluation protocol support further study of force estimation and physical interaction.

AI use statement

We used large language models to polish the manuscript and assist with the literature review.

REFERENCES

Ademi Adeniji, Zhuoran Chen, Vincent Liu, Venkatesh Pattabiraman, Raunaq Bhirangi, Siddhant Haldar, Pieter Abbeel, and Lerrel Pinto. Feel the force: Contact-driven learning from humans. *arXiv preprint arXiv:2506.01944*, 2025.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Robert Equi, Chelsea Finn, Niccolo Fusai, Manuel Y Galliker, et al. $\pi_{0.5}$: A vision-language-action model with open-world generalization. In *9th Annual Conference on Robot Learning*, 2025.

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Robotics: Science and Systems*, 2023.

Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shuran Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. In *Robotics: Science and Systems (RSS)*, 2024.

Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, et al. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023.

Patrick Grady, Chengcheng Tang, Samarth Brahmbhatt, Christopher D Twigg, Chengde Wan, James Hays, and Charles C Kemp. Pressurevision: estimating hand pressure from a single rgb image. In *European Conference on Computer Vision*, pp. 328–345. Springer, 2022.

Patrick Grady, Jeremy A Collins, Chengcheng Tang, Christopher D Twigg, Kunal Aneja, James Hays, and Charles C Kemp. Pressurevision++: Estimating fingertip pressure from diverse rgb images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 8698–8708, 2024.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3, 000 hours of egocentric video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 18973–18990. IEEE, 2022.

Konstantin Gubernatorov, Mikhail Sannikov, Ilya Mikhalchuk, Egor Kuznetsov, Makar Artemov, Ogunwoye Faith Ouwatobi, Marcelino Fernando, Artem Asanov, Ziang Guo, and Dzmitry Tsetserukou. HapticVLA: Contact-rich manipulation via vision-language-action model without inference-time tactile sensing. *arXiv preprint arXiv:2603.15257*, 2026.

Irmak Guzey, Yinlong Dai, Ben Evans, Soumith Chintala, and Lerrel Pinto. See to touch: Learning tactile dexterity through visual incentives. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 13825–13832. IEEE, 2024.

Xuhui Kang, Tongxuan Tian, Sung-Wook Lee, Binghao Huang, Yunzhu Li, and Yen-Ling Kuo. Learning force-regulated manipulation with a low-cost tactile-force-controlled gripper. *arXiv preprint arXiv:2602.10013*, 2026.

Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. Egomimic: Scaling imitation learning via egocentric video. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 13226–13233. IEEE, 2025.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. In *Conference on Robot Learning*, pp. 2679–2713, 2024.

Yang Li, Zhaxizhuoma, Hongru Jiang, Junjie Xia, Hongquan Zhang, Jinda Du, Yunsong Zhou, Jia Zeng, Ce Hao, Jieji Ren, Qiaojun Yu, Cewu Lu, Yu Qiao, and Jiangmiao Pang. ForceVLA2: Unleashing hybrid force-position control with force awareness for contact-rich manipulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Fangchen Liu, Chuanyu Li, Yihua Qin, Jing Xu, Pieter Abbeel, and Rui Chen. ViTaMIn: Learning contact-rich tasks through robot-free visuo-tactile manipulation interface. *arXiv preprint arXiv:2504.06156*, 2025.

Vimal Mollyn and Chris Harrison. Egotouch: On-body touch input using ar/vr headset cameras. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–11, 2024.

Yaru Niu, Zhenlong Fang, Binghong Chen, Shuai Zhou, Revanth Senthilkumaran, Hao Zhang, Bingqing Chen, Chen Qiu, H Eric Tseng, Jonathan Francis, et al. Learning versatile humanoid manipulation with touch dreaming. *arXiv preprint arXiv:2604.13015*, 2026.

Nicola A. Piga and Lorenzo Natale. Adaptive tactile force control in a parallel gripper with low positioning resolution. *IEEE Robotics and Automation Letters*, 8(9):5544–5551, 2023. doi: 10.1109/LRA.2023.3297061.

Carmelo Sferrazza, Younggyo Seo, Hao Liu, Youngwoon Lee, and Pieter Abbeel. The power of the senses: Generalizable manipulation from vision and touch through masked multimodal learning. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 9698–9705, 2024.

Siqi Shang, Minchao Huang, Bill Fan, and Lillian Chin. Exp-Force: Experience-conditioned pregrasp force selection with vision-language models. *arXiv preprint arXiv:2603.08668*, 2026.

Mustafa Shukor, Dana Aubakirova, Francesco Capuano, Pepijn Kooijmans, Steven Palma, Adil Zouitine, Michel Aractingi, Caroline Pascal, Martino Russi, Andres Marafioti, et al. Smolvla: A vision-language-action model for affordable and efficient robotics. *arXiv preprint arXiv:2506.01844*, 2025.

Yuxin Ray Song, Jinzhou Li, Rao Fu, Devin Murphy, Kaichen Zhou, Rishi Shiv, Yaqi Li, Haoyu Xiong, Crystal Elaine Owens, Yilun Du, et al. Opentouch: Bringing full-hand touch to real-world interaction. *arXiv preprint arXiv:2512.16842*, 2025.

Chen Wang, Haochen Shi, Weizhuo Wang, Ruohan Zhang, Li Fei-Fei, and C. Karen Liu. DexCap: Scalable and portable mocap data collection system for dexterous manipulation. In *Robotics: Science and Systems (RSS)*, 2024.

Songlin Wei, Hongyi Jing, Boqian Li, Zhenyu Zhao, Jiageng Mao, Zhenhao Ni, Sicheng He, Jie Liu, Xiawei Liu, Kaidi Kang, Sheng Zang, Weiduo Yuan, Marco Pavone, Di Huang, and Yue Wang. ψ_0: An open foundation model towards universal humanoid loco-manipulation. *arXiv preprint arXiv:2603.12263*, 2026.

Daniel M Wolpert and J Randall Flanagan. Motor prediction. *Current biology*, 11(18):R729–R732, 2001.

Daniel M. Wolpert, Zoubin Ghahramani, and Michael I. Jordan. An internal model for sensorimotor integration. *Science*, 269(5232):1880–1882, 1995.

Longyan Wu, Jieji Ren, Chenghang Jiang, Junxi Zhou, Shijia Peng, Ran Huang, Guoying Gu, Li Chen, and Hongyang Li. TAMEn: Tactile-aware manipulation engine for closed-loop data collection in contact-rich tasks. *arXiv preprint arXiv:2604.07335*, 2026a.

Qiwei Wu, Rui Zhang, Xin Xiang, Tao Li, Weihua Zhang, Junjie Lai, and Renjing Xu. Tabero: Learning gentle manipulation with closed-loop force feedback from vision, touch, and language. *arXiv preprint arXiv:2605.27886*, 2026b. URL `https://arxiv.org/abs/2605.27886`.

Kelin Yu, Yunhai Han, Qixian Wang, Vaibhav Saxena, Danfei Xu, and Ye Zhao. Mimictouch: Leveraging multi-modal human tactile demonstrations for contact-rich manipulation. In *Conference on Robot Learning*, pp. 4844–4865. PMLR, 2024.

Andy Zhai, Brae Liu, Bruno Fang, Chalse Cai, Ellie Ma, Ethan Yin, Hao Wang, Hugo Zhou, James Wang, Lights Shi, Lucy Liang, Make Wang, Qian Wang, Roy Gan, Ryan Yu, Shalfun Li, Starrick Liu, Sylas Chen, Vincent Chen, and Zach Xu. Igniting vlms toward the embodied space. *arXiv preprint arXiv:2509.11766*, 2025.

Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Robotics: Science and Systems*, 2023.

Yiming Zhao, Taein Kwon, Paul Streli, Marc Pollefeys, and Christian Holz. EgoPressure: A dataset for hand pressure and pose estimation in egocentric vision. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27727–27738, 2025.

Jinliang Zheng, Jianxiong Li, Zhihao Wang, Dongxiu Liu, Xirui Kang, Yuchun Feng, Yinan Zheng, Jiayin Zou, Yilun Chen, Jia Zeng, et al. X-vla: Soft-prompted transformer as scalable cross-embodiment vision-language-action model. *arXiv preprint arXiv:2510.10274*, 2025.

Yuhang Zheng, Songen Gu, Weize Li, Yupeng Zheng, Yujie Zang, Shuai Tian, Xiang Li, Ce Hao, Chen Gao, Si Liu, Haoran Li, Yilun Chen, Shuicheng Yan, and Wenchao Ding. OmniVTA: Visuo-tactile world modeling for contact-rich robotic manipulation. *arXiv preprint arXiv:2603.19201*, 2026.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pp. 2165–2183. PMLR, 2023.

A Reproducible Hardware Blueprint and Synchronization

To support replication of controlled force-intuition probes, we release the hardware blueprint and synchronization pipeline for the FEEL collection setup.

A.1 Bill of Materials (BOM)

Our design prioritizes low cost and accessibility. The total cost of the tactile enhancement (excluding the Meta Aria glasses) is kept under $500 USD.

Table 2: Bill of Materials for the FEEL egocentric force-probe collection setup.

Component	Description	Approx. Cost (USD)
Meta Aria Glasses	Egocentric vision and 6-DoF MPS tracking	N/A (Research Program)
Tactile Sensor	Paxini PX6AX-GEN3-DP-S2015-Elite ($15 \times 20 \times 10$ mm)	$150
3D Printed Mounts	Custom brackets for index finger and thumb (PLA/PETG)	$5
Microcontroller/Hub	For sensor data aggregation and timestamping	$35
Cables & Straps	Velcro straps, USB-C cables, and routing clips	$10

A.2 Spatio-Temporal Synchronization Pipeline

Aligning the tactile stream (50 Hz) with the Meta Aria RGB stream (10 Hz) is critical for capturing rapid force transients like force overshoot. We utilize a software-based synchronization approach:

1. **Clock Synchronization**: The tactile sensor hub and the Aria companion app are synchronized to a local NTP server before data collection.
2. **Event-based Alignment**: A physical tapping gesture at the beginning of each trajectory produces a sharp spike in the tactile force reading, which is cross-correlated with the Aria IMU acceleration spike to compute a temporal offset Δt.
3. **Spatial Calibration**: The tactile sensor's spatial frame is calibrated to the Aria wrist coordinate frame using a standard ArUco marker board.

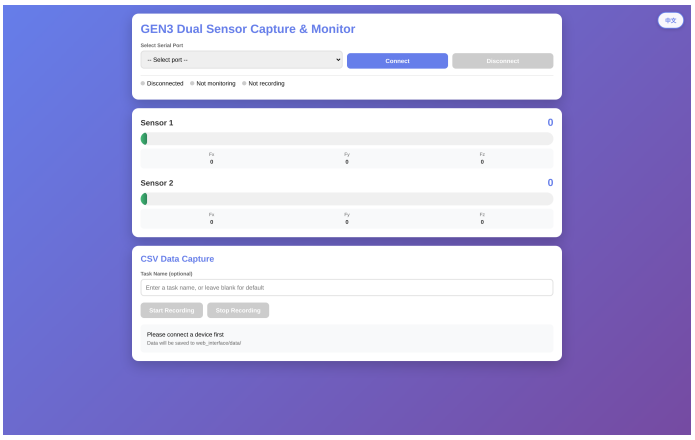
A.3 Annotation & Synchronization Web UI

To make the FEEL data engine usable beyond the original collection setup, we release a browser-based UI that wraps the three operations a new user must perform: (i) live tactile capture and CSV recording (Fig. 11a), (ii) post-hoc temporal alignment between the recorded tactile stream and the Aria visual/MPS streams (Fig. 12), and (iii) episode-level data management and quality inspection (Fig. 11b). The browser interface supports tactile CSV capture, aligned LeRobot episode export, and inspection of the resulting dataset.

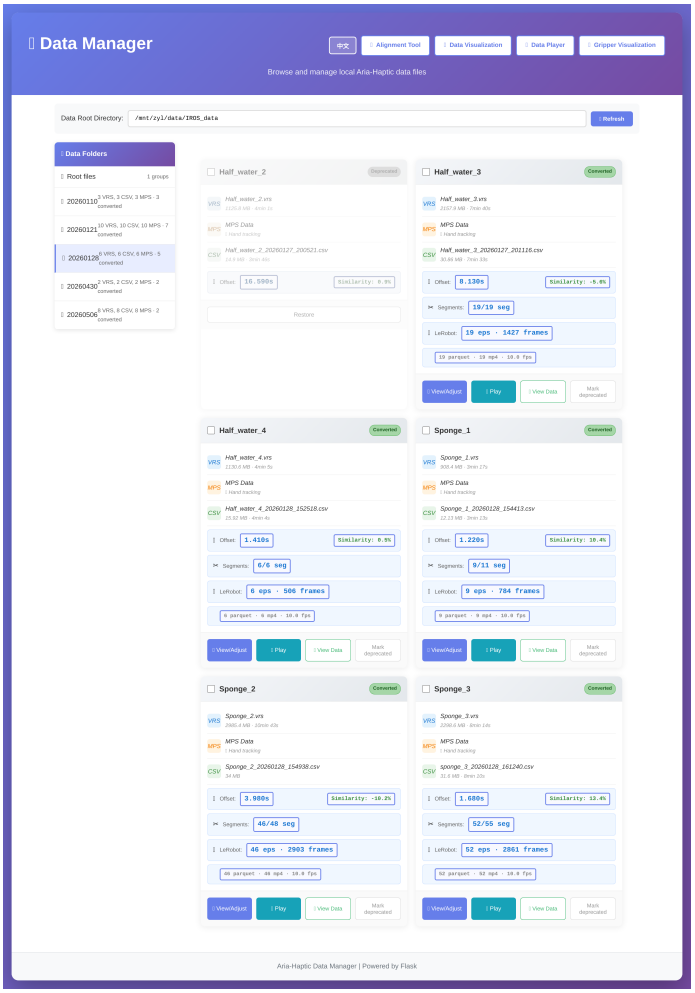
Capture module (Fig. 11a). The capture page enumerates connected GEN3 / Paxini sensors, displays F_x, F_y, F_z per sensor at the native sampling rate, and writes timestamped CSV streams to the LeRobot dataset layout. A task-name field tags the resulting episode for downstream stratification.

Time-alignment module (Fig. 12). The alignment page consumes the recorded haptic CSV, the MPS hand-tracking CSV, and the Aria VRS file, then estimates the temporal offset between the streams. The annotator can override the auto-detected offset via a slider, inspect the synchronized RGB-vs-force timeline, edit force segments, and export aligned episodes in LeRobot format with a single click. Sample latency, mean / median latency, and a quality score are reported per session for quality assurance, supporting inspection of temporal alignment between the recorded streams.

Data-manager module (Fig. 11b). The data manager indexes local Aria-haptic episodes, groups them by collection folder (e.g., Half_water, Sponge), and surfaces per-episode metadata—MPS / haptic / video presence, alignment offset, duration, accuracy, frame count—together with playback and visualization shortcuts. This view supports stratified split construction and dataset quality assurance by combining episode metadata with playback and visualization of the recorded streams.



(a) Capture & Monitor.



(b) Data Manager.

Figure 11: The FEEL annotation web UI: capture and data management. (a) Dual-sensor live capture with per-axis force readouts (F_x, F_y, F_z) and one-click CSV recording into a LeRobot-compatible episode directory. (b) Data manager: browses local Aria-haptic episodes by folder, shows MPS / haptic / video status, alignment offset, duration, sample-accuracy, and frame counts per episode, with one-click playback and visualization.

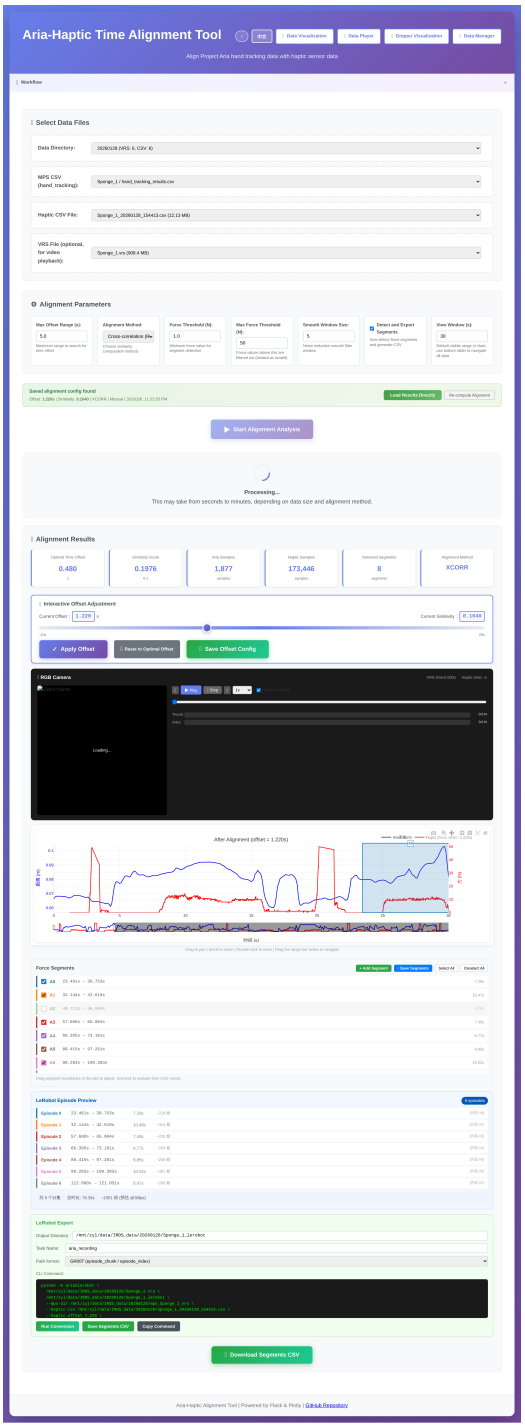


Figure 12: The FEEL Aria–Haptic time-alignment tool: loads MPS hand-tracking, haptic CSV, and Aria VRS files; runs cross-correlation alignment with adjustable thresholds; and exposes an interactive offset slider, force-segment editor, and per-episode LeRobot export pipeline.

The project page provides access to dataset resources and collection demonstrations (`http://submission.anonymoussubmission.com/`).

B Implementation Details and Hyperparameters

B.1 EgoForce Architecture Details

EgoForce uses an ImageNet-initialized ResNet-18, a four-layer Transformer encoder, and a one-layer decoder with model dimension 512, eight attention heads, and feed-forward dimension 3200. The encoder combines egocentric RGB with a 7D hand state. A GRU with hidden size 512 processes four consecutive observations; its hidden state is reset for each observation window. Each finger-specific shared head has dimensions $512 \to 256 \to 2$ with GELU and outputs a base force mean $\mu_{t,f}^{\text{base}}$ and log scale $s_{t,f}$, both defined in the quantile-normalized force space used for training.

A linear $512 \to 2$ router computes softmax weights from the final GRU hidden state. Two residual MLPs, each $512 \to 128 \to 2$ with GELU, produce the mean correction in Equation 2. Their hidden layers are independently initialized and their final projections are initialized to zero; both experts are active for every sample. Scale predictions remain in the shared heads. The force objective uses $\bar{s} = \text{clip}(s, -5, 2)$ in both the scale-weighted absolute error and the additive log-scale penalty, as in Equation 1. The model has 36 177 101 parameters, compared with 36 044 231 for the shared-head model; the additional 132 870 parameters belong to the two residual MLPs and linear router.

Training uses AdamW, learning rate 10^{-5} for the backbone and remaining layers, weight decay 10^{-4}, gradient clipping at 1.0, batch size 4, and 10 000 steps with seed 1000. Images are resized to 480×480; state and force targets use quantile normalization. The contact auxiliary branch is $512 \to 64 \to 1$ with GELU and sigmoid. Its binary target is $\mathbb{1}[\max_f F_{t,f}^{\text{gt}} > 0\,\text{N}]$, computed from measured force before quantile normalization, and its BCE weight is 0.1. The general expert implementation also includes a load-balance term with weight 0.01; for the two-expert configuration with both experts active, this term is constant. Router weights learn through the force objective. Evaluation applies inverse normalization, clips negative force predictions to zero, and uses a 0 N contact threshold. Training of the reported model took 35.21 minutes on one A100 during concurrent experiments.

B.2 Evaluated Architectures and Force-Output Adaptation

All eight baselines use RGB and 7-D hand-state observations to predict two fingertip forces. **ACT** uses a Transformer encoder–decoder with direct force regression. The **OpenTouch adaptation** freezes a DINOv2 ViT-B/14 backbone and combines its visual features with a pose MLP ($7 \to 128 \to 256 \to 256$), followed by a regression head ($1024 \to 512 \to 256 \to 2$). The six **flow-matching VLAs**—Pi0, Pi0.5, WALL-X, SmolVLA, X-VLA, and GR00T—adapt their action outputs to the same force targets using model-specific padding and normalization. Their backbones differ: X-VLA uses Florence-2 with a soft-prompted action Transformer, while GR00T uses an Eagle VLM and a DiT action head trained with flow matching. GR00T is trained with HuggingFace Trainer. Measured tactile forces provide supervision targets.

B.3 Per-Baseline Training Configuration

The main benchmark uses 10 000-step checkpoints trained on FEEL (1 937 training episodes, 55 525 frames), with seed 1000. Table 3 summarizes the recorded training configurations. Applicable baselines are evaluated at chunk size 1 and their default chunk sizes. Both EgoForce variants and ACT use ImageNet-initialized ResNet-18 backbones. EgoForce uses ImageNet backbone initialization and newly initialized task modules.

Table 4 summarizes the model architectures and parameter counts.

B.4 Evaluation Details

For contact Pearson r, eligible constant predictions receive zero. The minimum sample count and ground-truth variation requirements in Section 5 apply only to this trend metric. Ground-truth eligibility covers 1 697/1 742 finger–episode pairs (97.42%) on Test and 508/522 (97.32%) on OOD. Other undefined metric values are excluded from the corresponding aggregates. Predictions are evaluated without temporal smoothing or alignment.

Table 3: Training configurations for the main benchmark. Batch denotes effective batch size; GR00T uses batch size 1 with 32-step gradient accumulation. LR is the configured learning rate.

Model	Batch	Chunk sizes	LR	Weight decay
EgoForce (MoE / w/o MoE)	4	1	$1e{-5}$	$1e{-4}$
ACT	4	1, 100	$1e{-5}$	$1e{-4}$
WALL-X	4	1, 32	$2e{-5}$	$1e{-2}$
Pi0	8	1, 50	$2.5e{-5}$	$1e{-2}$
Pi0.5	32	1, 50	$2.5e{-5}$	$1e{-2}$
X-VLA	4	1, 30	$2.5e{-5}$	0
SmolVLA	8	1, 50	$2.5e{-5}$	$1e{-10}$
OpenTouch	32	1	$3e{-4}$	$1e{-4}$
GR00T	32	1, 16	$1e{-5}$	$1e{-5}$

Table 4: Backbone, total parameter count, and trainable parameter count per model.

Model	Backbone / architecture	Total params	Trainable params
EgoForce (MoE)	GRU + shared Laplace heads + 2 residual experts	36.18 M	36.18 M
EgoForce w/o MoE	GRU + shared Laplace heads	36.04 M	36.04 M
ACT	Transformer encoder–decoder	∼34 M	∼34 M
WALL-X	Qwen2.5-VL + flow matching	∼4.2 B	∼4.2 B
Pi0	PaliGemma + action expert	∼3.5 B	∼578 M
Pi0.5	PaliGemma + action expert (AdaRMS)	∼3.6 B	∼693 M
X-VLA	Florence-2 + flow-matching action Transformer	∼880 M	∼880 M
SmolVLA	SmolVLM2-500M-Video + flow matching	∼460 M	∼100 M
OpenTouch	DINOv2 ViT-B/14 + pose MLP + regression head	∼87 M	∼756 K
GR00T	Eagle VLM + flow-matching DiT head	∼3 B	∼751 M

B.5 Training Dynamics

Figure 13 shows the total training losses for the main-benchmark runs evaluated in Table 1: seed 1000 and 10 000 steps, including EgoForce, its shared-head variant, and all eight baselines. Solid curves use chunk size 1; dashed curves show each baseline's default chunk size or the shared-head variant in the EgoForce panel. Each point averages the training loss over a successive 500-step window. The source logs record 20-step means for both EgoForce variants, 500-step means for the LeRobot baselines, and 10-step means for GR00T; the shorter intervals are aggregated to the common window. Each panel retains its model's objective scale. For EgoForce and its shared-head variant, the logged loss includes the force objective and weighted contact BCE; EgoForce also includes a constant 0.02 load-balance term.

The first-to-final window loss decreases in every run. For EgoForce and its shared-head variant, the total training losses decrease from -0.024 to -1.225 and from -0.064 to -1.223, respectively. Later changes vary across models: at chunk size 1, the mean loss over steps 8 001–9 000 versus 9 001–10 000 changes from 0.2505 to 0.2475 for Pi0, from 0.1460 to 0.1455 for OpenTouch, and from 0.3020 to 0.2850 for SmolVLA. EgoForce's training loss also decreases over these windows, from -1.152 to -1.203. The curves characterize optimization under the common training budget; the held-out force metrics are reported in Table 1.

C Additional Ablation Results

To examine the architectural choices in **Q1**, we compare EgoForce with three variants that separately remove the residual experts, replace the GRU with temporal pooling, or omit hand-state information. All use the main-benchmark splits, seed 1000, batch size 4, and final 10 000-step checkpoints, with the same bounded Laplace objective and contact auxiliary supervision.

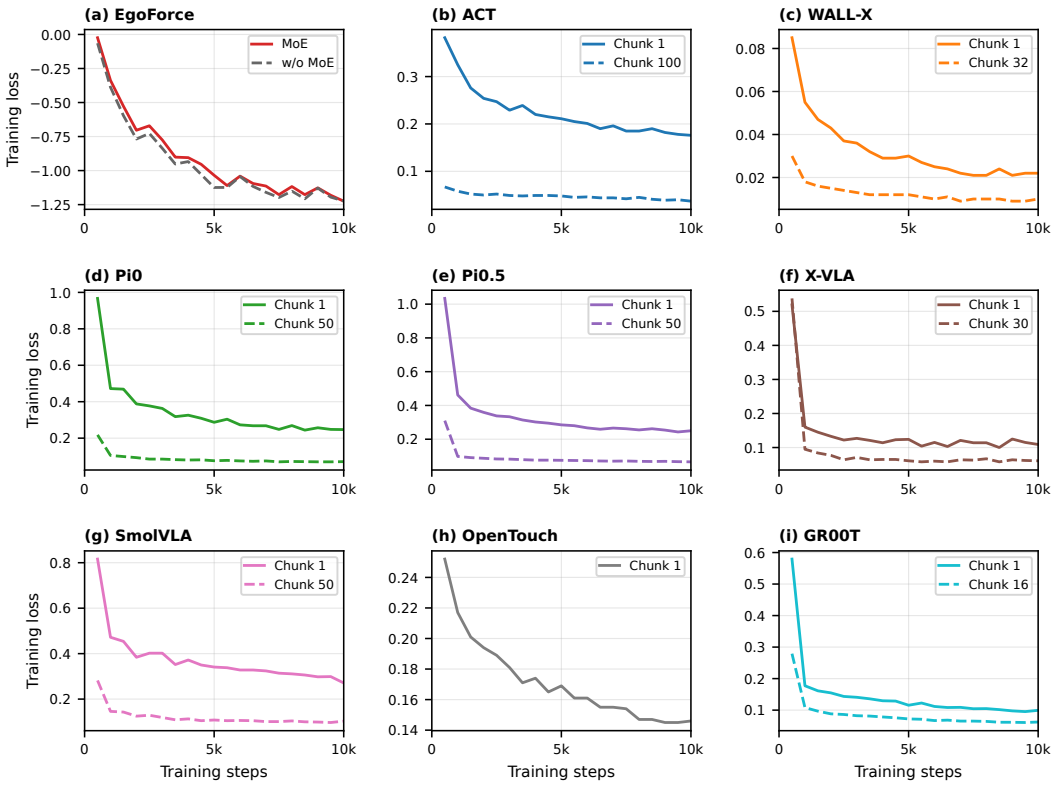


Figure 13: Training dynamics of the main benchmark runs, seed 1000. Points are means over successive 500-step windows in the original total-loss units; panels have independent vertical scales. Solid: chunk size 1. Dashed: default chunks, or EgoForce w/o MoE in panel (a).

C.1 Expert, Temporal, and State Ablations

The shared-head variant removes only the residual experts and their router. The temporal variant replaces the GRU with an average of the encoded features over the same four observations, excluding padded frames. The state variant sets the normalized hand-state input to zero while retaining the state encoder. Table 5 reports all four metrics on Test and OOD.

Table 5: Component ablations on FEEL, seed 1000 and 10 000 training steps. Each variant changes the named component while retaining the remaining training settings. Contact MAE is in N. Bold denotes the best value within each split.

Variant	MAE↓	R^2 ↑	Pearson r ↑	Pooled F1↑
Test: contact metrics				
EgoForce	7.3386	0.1965	**0.3022**	0.8395
w/o MoE	7.3905	0.1941	0.2963	**0.8600**
w/o GRU (four-frame pooling)	**7.1933**	**0.2333**	0.2449	0.8417
w/o hand state	7.4044	0.1817	0.2995	0.8424
OOD: contact metrics				
EgoForce	4.3963	**-0.7672**	**0.4067**	**0.8934**
w/o MoE	6.9350	-5.4770	0.3920	0.8674
w/o GRU (four-frame pooling)	4.6238	-0.8118	0.3294	0.8483
w/o hand state	**4.1392**	-1.3836	0.3967	0.8487

Residual experts. Adding the expert branch improves Test contact MAE from 7.3905 to 7.3386 N, contact R^2 from 0.1941 to 0.1965, and contact r from 0.2963 to 0.3022; pooled F1 changes from 0.8600 to 0.8395. On OOD, all four metrics improve: MAE decreases by 36.61%, contact R^2 rises from -5.4770 to -0.7672, contact r from 0.3920 to 0.4067, and pooled F1 from 0.8674 to 0.8934. The residual branch therefore provides its largest magnitude improvement on the held-out container condition.

Temporal aggregation. The GRU improves all four OOD metrics over four-frame pooling, including contact r from 0.3294 to 0.4067 and pooled F1 from 0.8483 to 0.8934. On Test, contact r increases from 0.2449 to 0.3022, while pooling gives lower MAE and higher R^2 and F1. These comparisons associate recurrent aggregation with stronger within-episode force correlation and OOD performance.

Hand-state input. Including the hand state improves contact R^2 and r on both splits. On Test, MAE decreases from 7.4044 to 7.3386 N, while F1 changes from 0.8424 to 0.8395. On OOD, pooled F1 increases from 0.8487 to 0.8934, alongside improved R^2 and r; the state-free variant has lower MAE (4.1392 vs. 4.3963 N). The state input thus contributes to correlation and contact detection under the held-out condition.

C.2 Additional Visualizations

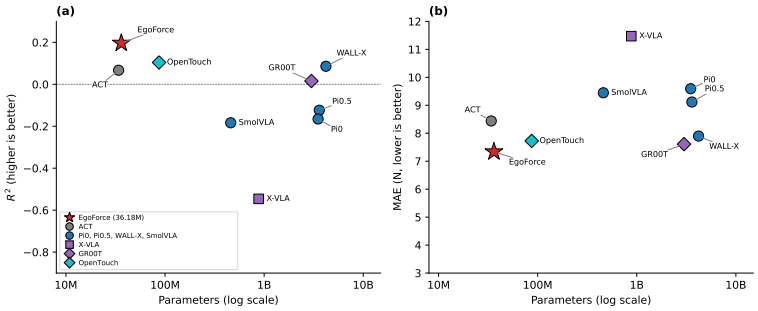


Figure 14: Total parameter count versus Test contact performance, seed 1000 and chunk size 1. EgoForce has 36.18 M parameters. Colors and marker shapes distinguish the listed models and model groups. The panels use the contact R^2 and contact MAE definitions of Table 1.

D Real-Gripper Force Tracking Details

D.1 Hardware and Feedback

The platform uses an Alicia-M arm carrying a gripper with one opening actuator and two electronic-skin pads. The experiment adjusts the gripper opening while maintaining the arm configuration. Each pad provides 60 scalar readings $p_{i,j}$ in mN. The per-side forces and the feedback used by the controller are

$$F_i = 10^{-3} \sum_{j=1}^{60} p_{i,j}, \qquad F_{\mathrm{fb}} = \frac{F_L + F_R}{2}, \tag{3}$$

where F_i and F_{fb} are expressed in N. All sensing points contribute to the feedback, without threshold filtering or software tare. Force values use the sensor's protocol conversion; independent absolute-force calibration was not performed for this run.

D.2 Controller and Protocol

We adapt the controller of Piga & Natale (2023), retaining its local stiffness estimation and quantization-aware position update. After contact, the base update is $\Delta q = -K(F_{\mathrm{ref}} - F_{\mathrm{fb}})\Delta t$, subject to position and speed limits and the stiffness-based update check. Decreasing the motor

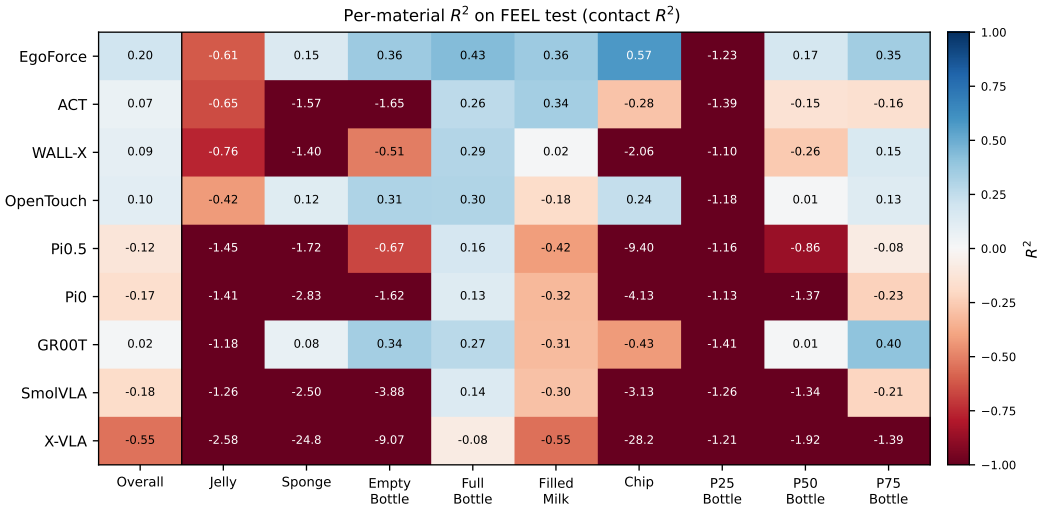


Figure 15: Per-material contact R^2 on the FEEL test split, seed 1000 and chunk size 1. Contact frames are pooled per finger within each category, then the two scores are averaged. Colors are clipped to $[-1, 1]$; cell text gives the actual value.

angle closes the gripper. The internal force reference approaches the prescribed target at 0.5 N/s. Contact detection uses the same mean-force feedback; the position reference is brought to the measured contact position at a limited rate before regular force adjustment. A motor-torque saturation check prevents further integration toward the configured limit.

Table 6: Fixed settings for the real-gripper experiment.

Parameter	Value
Force-error gain K	0.05 rad/(N·s)
Position-reference speed limit	0.03 rad/s
Force-reference rate	0.5 N/s
Contact threshold	0.05 N
Closed/open motor-angle limits	0.003/2.426 rad
Configured holding-torque limit	2.5 N·m
Evaluation interval per target	60 s
Fixed final-window duration	10 s
Settling band and minimum dwell	± 0.05 N; 10 s

The object is a tissue package. Initial grasp establishment ends after the 1 N feedback remains within ± 0.05 N for at least 10 s. A single continuous sequence then evaluates targets $1, 2, 3, 4, 5, 4, 3, 2, 1$ N using one controller instance. The first interval is a 1 N holding baseline. Each evaluation window spans the first 60 s after the corresponding target is applied; target changes include a short communication interval to finish recording in-flight measurements. The acquisition rate is approximately 4.54 Hz. Raw tactile samples are paired with their corresponding control updates in the backend log.

D.3 Metrics and Results

For a target $F^\star$, tracking MAE is the arithmetic mean of $|F_{\text{fb}} - F^\star|$ over the 2 457 samples in the nine 60 s evaluation windows. The final-window MAE pools the 412 samples in the fixed last 10 s of these windows. These intervals are specified before execution. The corresponding MAEs are 0.0487 N and 0.0207 N; the final-window RMSE is 0.0291 N, and 92.7% of final-window samples lie within ± 0.05 N of the target.

Settling time starts when the target is applied and ends at the first sample of the final uninterrupted in-band segment, which must last at least 10 s through the end of the evaluation window. Scoring requires observations within 0.8 s of both interval boundaries and no internal sampling gap above

0.8 s. Four intervals, including the baseline, satisfy the terminal dwell criterion. Increasing-target overshoot is $\max(0, \max_t F_{\text{fb}} - F^{\star})$; decreasing-target undershoot is $\max(0, F^{\star} - \min_t F_{\text{fb}})$. All curves and statistics use recorded samples without smoothing or resampling.

Table 7: Per-stage results. Force means, MAE, and peak deviations are in N. Final-window statistics use the fixed last 10 s. Peak deviation denotes overshoot on ascent and undershoot on descent. A dash in settling time denotes either an interval that did not meet the terminal dwell criterion or the initial 1 N holding baseline, which uses a constant target before the sequence of target changes.

Target change	Final-window mean	Final-window MAE	In band (%)	Peak deviation	Settling time (s)
1 (baseline)	1.0101	0.0101	100.0	—	—
$1 \rightarrow 2$	2.0007	0.0143	100.0	0.3165	5.05
$2 \rightarrow 3$	2.9866	0.0134	100.0	0.2855	5.62
$3 \rightarrow 4$	4.0001	0.0062	100.0	0.7680	9.18
$4 \rightarrow 5$	5.0004	0.0186	95.7	0.2790	—
$5 \rightarrow 4$	3.9688	0.0312	97.7	0.1385	—
$4 \rightarrow 3$	3.0004	0.0179	95.7	0.0910	—
$3 \rightarrow 2$	2.0255	0.0481	60.5	0.2560	—
$2 \rightarrow 1$	0.9998	0.0308	81.4	0.0860	—

The largest increasing-target peak occurs during the $3 \rightarrow 4$ N transition, reaching 4.768 N. At the descending 2 N target, the final-window MAE is 0.0481 N and the in-band fraction is 60.5%, compared with 0.0143 N and 100% for the ascending 2 N target. These observations separate transient peaks from the later fluctuations visible in Figure 9.

E Simulation: Force-Action Perturbations

We evaluate the closed-loop Tabero-VTLA policy (Wu et al., 2026b) on the Milk task in Isaac Sim. The policy supplies force and position commands to a hybrid controller. The intervention perturbs the predicted bilateral force-action channels, with initial states and sampling noise paired across perturbation levels. The sweep spans injected force-action MAE from 0 to 20 N in 2 N increments, with 150 trials at each level. Figure 16 illustrates the closed-loop execution sequence, from approaching and grasping the milk container through lifting, transport, and placement at the destination.

Figure 8 reports pooled success rates and force-tracking MAE. Success decreases from 84.7% without injected perturbations to 53.3% at 4 N and 28.0% at 6 N, reaching zero at 16–20 N. Force-tracking MAE increases from 10.4 N at the zero-perturbation level to 21.2 N at 20 N. The success panel also displays individual random-number-generation blocks and 95% episode-cluster bootstrap confidence intervals. These measurements characterize how errors in the commanded forces propagate to tracking and task completion in this closed-loop policy.



Figure 16: A Tabero-VTLA Milk rollout in Isaac Sim. The sequence shows the robot approaching and grasping the milk container, lifting and transporting it, and placing it at the destination.

F LIMITATIONS AND BROADER IMPACTS

Limitations: FEEL evaluates offline force estimation from four demonstrators, while the robot experiment measures tracking of prescribed force targets. Cross-person generalization and transfer of predicted forces to robot control remain open directions. The benchmark reports seed-1000 results; repeated training runs and broader OOD categories can characterize variability across optimization and object conditions. Longer collection periods also call for studies of tactile calibration consistency and sensor wear, including how force measurements change with repeated sensor use.

Broader Impacts: The FEEL collection setup and force-probe protocol provide resources for studying physical priors in egocentric learning. Better force estimates can support more precise manipulation of fragile and deformable objects. Deployment studies should evaluate force limits, calibration, and behavior under unfamiliar objects and sensor conditions.